Reeve, H., & Kaban, A. (2019). Robust randomized optimization with k nearest neighbors. *Analysis and Applications*, *17*(5), 819-836. https://doi.org/10.1142/S0219530519400086

Analysis and Applications

Robust randomised optimisation with k nearest neighbours

Henry W J Reeve
Ata Kabán
School of Computer Science,
University of Birmingham,
B15 2TT, UK

Received 15th April 2019

Modern applications of machine learning typically require the tuning of a multitude of hyper-parameters. With this motivation in mind, we consider the problem of optimisation given a set of noisy function evaluations. We focus on robust optimisation in which the goal is to find a point in the input space such that the function remains high when perturbed by an adversary within a given radius. Here we identify the minimax optimal rate for this problem, which turns out to be of order $\mathcal{O}(n^{-\lambda/(2\lambda+1)})$, where n is the sample size and λ quantifies the smoothness of the function for a broad class of problems, including situations where the metric space is unbounded. The optimal rate is achieved (up to logarithmic factors) by a conceptually simple algorithm based on k-nearest neighbour regression.

Keywords: Optimisation for machine learning; metric spaces; non-parametric methods.

Mathematics Subject Classification 2010: 68Txx, 62Gxx, 62Pxx

1. Introduction

Machine learning algorithms typically require the user to tune a multiplicity of hyper-parameters. For example, an SVM with Gaussian kernel requires the user to select a regularisation penalty and a Gaussian bandwidth. This problem has been exacerbated by the current trend towards complex deep neural networks which possess a diverse assortment of different hyper-parameters which characterise the architecture (number of layers, nodes per layer, kernel size etc.), the regularisation settings (dropout, weight decay, data augmentation etc.) and the training procedure (weight initialisation, learning rate, annealing schedule etc.) [13,22,28,31]. A configuration of hyper-parameters may be viewed as point in a metric space, with its performance estimated by evaluating the corresponding trained model on a validation set [4,5,13]. However, the performance of a configuration on a particular randomly selected validation set is only a proxy for the true value of interest - performance on unseen data from the underlying distribution. Hence, hyper-parameter tuning based on validation performance may be viewed as an optimisation problem with noisy function evaluations [5,9,24,25].

The problem of hyper-parameter tuning is complicated by the potential for mismatch between the validation data we have access to and the underlying distribution that the trained model will ultimately be used upon [29]. This motivated Bogunovic et al [9] to model hyper-parameter tuning as a problem of *robust optimisation*, in which the goal is to find a point in the input space such that the function remains high when perturbed by an adversary within a given radius. Further motivation for robust optimisation comes from the observation that broad flat optima typically generalise better than sharp narrow optima in the context of overparameterised models [1,16]. The problem of robust optimisation is also natural in other situations where there is a mismatch between the noisy function evaluations we have access to and the true function. For example, in robotics, functions are often optimised using simulation data rather than measurements of the physical robot. Robust optimisation is also applicable to situations where we seek to find a region of good points in the input space, rather than a single point.

We shall consider the setting of randomised optimisation whereby a large number of input points are selected at random and noisy function evaluations for those input points are computed in parallel [5]. The major advantage of this approach is the reduction in running time achieved by performing function evaluations in parallel. This advantage is of paramount importance for applications to training complex models on large data sets. In contrast a sequential optimisation procedure would require a far greater run time, without utilising a distributed architecture. A secondary benefit of the randomised approach is that the procedure does not to be re-run when the performance metric is modified.

In this setting we shall consider a conceptually simple k-nearest neighbour based method for both optimisation and robust optimisation.

Contributions: We make the following contributions:

- We consider a simple method for randomised function optimisation based on k-nearest neighbours and prove a high probability bound on the simple regret (Theorem 4.1). The high-probability bound depends purely upon smoothness properties of the function and does not depend upon the ambient dimension of the space or the VC dimension of the set of balls.
- We introduce an approach for performing adverserially robust optimisation based upon k-nearest neighbours. The method is conceptually simple and straightforward to implement. Our main result (Theorem 5.1) gives a high probability upper bound on the performance of this method. A key feature of the bound in Theorem 5.1 is that it is independent of the ambient dimension of the feature space.
- We provide a complementary lower bound (Theorem 6.1) which demonstrates that the upper bounds given by Theorems 4.1 and 5.1 are optimal up to logarithmic factors.

Related work: The problem of robust optimisation has received substantial attention in the literature (see [3,2,7,6,8] and references within). The primary difference between the present paper and these works is that we do not assume that the function to be optimised is known by the learner; our only access to the underlying function is via noisy function evaluations. The most similar work to ours is the recent work of Bogunovic et al [9], where optimisation is performed based on noisy function evaluations with a Gaussian process based approach. In our work we consider the randomised parallel setting, whereas [9] considers the sequential setting. Moreover, Bogunovic et al assume that the underlying function to be optimised is itself a member of the reproducing kernel Hilbert space for some known kernel. On the other hand, the classes considered in this paper are specified purely in terms of the level of smoothness of the underlying function.

Jiang [17] recently gave related bounds on the k-nearest neighbour method for function maximisation. However, [17] focuses on the distance in input space, whereas we focus on the deviation in function value (more natural in the optimisation setting). Note that studying the distance in input space requires strong second order assumptions not required in our setting. Note also that the constant terms in the bounds of [17] depend upon the ambient dimension of the space. This is a consequence of the fact that the analysis hinges upon a supremum norm bound, which we avoid in our approach. In a related work by the present authors [27] a bound has been established for determining the maximum value of a function. In this work we address the related question of determining an input point which attains near maximal value, before going on to consider the question of robust optimisation.

Our work draws upon the rich literature on k-nearest neighbour regression. In particular, we employ smoothness concepts from Chaudhuri and Dasgupta [10] and Kpotufe [20]. A related line of work in the literature are the Bayes consistent 1-nearest neighbour methods [14,15,18,19]. These works are focused upon classification, rather than optimisation, and as such do not make the smoothness assumptions required by [10] and the present work. However, some form of smoothness is vital for regret bounds in optimisation to preclude situations where the function to be opmised is jumps up on a set of zero measure. The analysis in the present paper leverages ideas from [20,10,21]. However, new ideas are required to deal with the challenge of robust optimisation in the non-parametric setting.

Outline: In Section 2 we precisly specify our problem setting. In Section 3 we introduce the k-nearest method. In Section 4 we consider the standard optimisation problem and give a simple k-nearest neighbour method with a high probabilty bound. In Section 5 we turn to the problem of robust optimisation and present our main result. In Section 6 we present our lower bound which demonstrates the optimality of the methods presented in Sections 4 and 5.

2. A statistical model for robust randomised optimisation

Suppose we have a metric space $(\mathcal{X}, \rho)$ along with an unknown function $f : \mathcal{X} \to [0,1]$. The goal of the learning agent is to select a point within the metric space which maximises f. We assume that the learning agent has access to a data set $\mathcal{D} = \{(X_i, Y_i)\}_{i\in[n]}$ with (X_i, Y_i) sampled i.i.d. from a distribution $\mathbb{P}$ on $\mathcal{X} \times \mathbb{R}$ satisfying $f(x) = \mathbb{E}_{\mathbb{P}}[Y|X = x]$. Throughout we let $[n] = \{1, \cdots, n\}$ and define the open and closed balls $B_r(x) = \{z \in \mathcal{X} : \rho(z, x) < r\}$ and $\overline{B}_r(x) = \{z \in \mathcal{X} : \rho(z, x) \leq r\}$. We also let μ denote the marginal distribution of $\mathbb{P}$ over $\mathcal{X}$ so $\mu(A) = \mathbb{P}[X \in A]$ for $A \subset \mathcal{X}$.

We shall begin by considering the problem of locating points $x(n)$ with function values close to the global maximum. To quantify our performance with respect to this problem we define the *regret* for $x \in \mathcal{X}$ by

$$\mathcal{R}_f(x) := \sup_{z\in\mathcal{X}} \{f(z)\} - f(x).$$

In addition we consider the problem of robust-optimisation. Take $\epsilon > 0$ and define for each $x \in \mathcal{X}$ the *adverserially peturbed* minimum $\underline{f}^{\epsilon}(x) = \inf_{z\in\overline{B}_{\epsilon}(x)}\{f(z)\}$ and define the *adverserially robust regret* by

$$\mathcal{R}^{\epsilon}_f(x) := \sup_{z\in\mathcal{X}} \left\{\underline{f}^{\epsilon}(z)\right\} - \underline{f}^{\epsilon}(x).$$

We shall make the following assumptions.

Assumption 2.1 (Noise assumption). *For all $x \in \mathcal{X}$, the deviation $Y - f(X)$ is a conditionally sub-Gaussian random variable with variance parameter σ^2 given $X = x$ i.e. $\forall t \in \mathbb{R}$, $\mathbb{E}[\exp(t \cdot (Y - f(X)))|X = x] \leq \exp(\sigma^2 \cdot t^2/2)$.*

Assumption 2.1 includes various natural examples including bounded and Gaussian noise.

Assumption 2.2 (Measure smoothness assumption). *Suppose that there are constants λ, $\omega > 0$ and $\epsilon > 0$ such that for all $x_0, x_1 \in \mathcal{X}$,*

$$|f(x_0) - f(x_1)| \leq \omega \cdot \inf_{z\in\overline{B}_{\epsilon}(x_0)} \left\{\mu\left(B_{\rho(x_0,x_1)}(z)\right)^{\lambda}\right\}.$$

Assumption 2.2 is a mild strengthening of the measure-theoretic smoothness condition of Chaudhuri and Dasgupta [10]. With $\epsilon = 0$ Assumption 2.2 corresponds to the measure-theoretic smoothness condition introduced in [10], which in turn generalises the modified Lipschitz condition of Döring et al [11]. In Proposition 2.1 we will give some readily checkable geometric properties which imply Assumption 2.2. However, we emphasise that the major advantage of Assumption 2.2 is that it does not require μ to be compactly supported and holds whenever f is sufficiently smooth in low-density regions of the space [10,11].

Proposition 2.1. *Suppose that f is α-Hölder continuous with constant $C_{\alpha} > 0$ for some $\alpha \in (0, 1]$ i.e. for all $x_0, x_1 \in \mathcal{X}$, $|f(x_0) - f(x_1)| \leq C_{\alpha} \cdot \rho(x_0, x_1)^{\alpha}$. Suppose*

further that μ is d-dimensional with constants $c_d > 0$, $d > 0$ i.e. for all $z \in \mathcal{X}$ and $r \in (0,1)$ we have $\mu(B_r(z)) \geq c_d \cdot r^d$. Then $\forall \epsilon > 0$, Assumption 2.2 holds with $\lambda = \alpha/d$ and $\omega = \max\{C_\alpha, 1\} \cdot c_d^{-\lambda}$.

The proof of Proposition 2.1 follows from the definitions and is given in Appendix Appendix C. When dealing with the robust optimisation problem we shall also require the following additional geometric property. In essence the property requires that between any two points we can draw a distance minimising curve.

Assumption 2.3 (Geodesic assumption). *Given any $x_0, x_1 \in \mathcal{X}$ there exists a $\rho(x_0, x_1)$-Lipschitz continuous map $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x_0$ and $\gamma(1) = x_1$.*

The following proposition gives two general settings in which Assumption 2.3 holds.

Proposition 2.2. *Suppose that $(\mathcal{X}, \rho)$ is either (a) a convex subset of a normed vector space or (b) a complete Riemannian manifold. Then Assumption 2.3 holds.*

Proof. In case (a) we can construct $\gamma : [0,1] \to \mathcal{X}$ with the desired properties by taking $\gamma(\theta) = x_0 + \theta \cdot (x_1 - x_0)$. It follows from convexity that $\gamma(\theta) \in \mathcal{X}$ for all $\theta \in [0,1]$. Moreover, it follows from the scalar property of the norm that γ is $\rho(x_0, x_1)$-Lipschitz. In case (b) we can take γ to be a minimising geodesic between x_0 and x_1. □

3. The k nearest neighbour method

Given $k \in [n]$ we define the k-nearest neighbour regression estimate $\hat{f}_{n,k}$ for the function f as follows. For each $x \in \mathcal{X}$ we let $\{\tau_q(x)\}_{q \in [n]}$ be a permutation of the indices $[n]$ so that for $q \in [n-1]$, $\rho(x, X_{\tau_q(x)}) \leq \rho(x, X_{\tau_{q+1}(x)})$. We then define $\hat{f}_{n,k} : \mathcal{X} \to \mathbb{R}$ by

$$\hat{f}_{n,k}(x) := \frac{1}{k} \sum_{q \in [k]} Y_{\tau_q(x)}.$$

Theorem 3.1. *Suppose that Assumptions 2.1 and 2.2 hold. Given $n \in \mathbb{N}$, $\delta \in (0,1)$, $k \in \{\lceil 4\log(2n/\delta) \rceil + 1, \cdots, \lfloor n/2 \rfloor\}$ the following holds with probability at least $1 - \delta$, for all $i \in [n]$,*

$$\left| \hat{f}_{n,k}(X_i) - f(X_i) \right| \leq \sigma \sqrt{\frac{2\log(4n/\delta)}{k}} + \omega \cdot \left(\frac{2k}{n} \right)^{\lambda}.$$

Note that Theorem 3.1 only becomes interesting when n is sufficiently large that $\{\lceil 4\log(2n/\delta) \rceil + 1, \cdots, \lfloor n/2 \rfloor\} \neq \emptyset$ (otherwise the result holds vacuously). The proof of Theorem 3.1 is standard and given in Appendix A for completeness.

4. Optimisation with k-NN

In this section we turn to the problem of maximising the function f given a data sample $\mathcal{D} = \{(X_i, Y_i)\}_{i\in[n]}$. Equivilantly, our goal is to choose a point $x \in \mathcal{X}$ which minimises the simple regret $\mathcal{R}_f(x)$. We propose the following simple estimator:

$$x_k(n) = \underset{X_i \,:\, i\in[n]}{\operatorname{argmax}} \left\{ \hat{f}_{n,k}(X_i) \right\}.$$

We simply perform k-nearest neighbour regression before computing the empirical maximum. The algorithm satisfies the following high probability bound on its performance.

Theorem 4.1. *Suppose that Assumptions 2.1 and 2.2 hold with $\epsilon = 0$. Given any $n \in \mathbb{N}$, $\delta \in (0,1)$ and $k \in \{\lceil 4\log(4n/\delta)\rceil + 1, \cdots, \lfloor n/2 \rfloor\}$ with probability at least $1-\delta$,*

$$\mathcal{R}_f\left(x_k(n)\right) \leq 2\sigma\sqrt{\frac{2\log(8n/\delta)}{k}} + 3\omega \cdot \left(\frac{2k}{n}\right)^{\lambda}.$$

In particular, with $k = \Theta\left((\log(n/\delta))^{1/(2\lambda+1)} \cdot n^{2\lambda/(2\lambda+1)}\right)$ then with probability at least $1-\delta$ we have

$$\mathcal{R}_f(x_k(n)) \leq \mathcal{O}\left(\left(\frac{\log(n/\delta)}{n}\right)^{\lambda/(2\lambda+1)}\right).$$

In Section 6 we will see that this is the best possible rate, up to logarithmic and constant factors. In this special case, optimality also follows from observations in the proof of [21, Theorem 1]. The proof of Theorem 4.1 is relatively straightforward but will provide a useful stepping stone towards the proof of Theorem 5.1.

Lemma 4.1. *Given any $n \in \mathbb{N}$, $\delta \in (0,1)$ and $x \in \mathcal{X}$ with probability at least $1-\delta$, $\mu(B_{\rho(x, X_{\tau_1(x)})}(x)) \leq \log(1/\delta)/n$.*

Proof. Choose $p = \log(1/\delta)/n$ and $r(p) = \inf\{r > 0 : \mu(B_r(x)) \geq p\}$, so $\mu(B_{r(p)}(x)) \leq p$ and $\mu(\overline{B}_{r(p)}(x)) \geq p$. It follows that $\{X_i\}_{i\in[n]} \cap \overline{B}_{r(p)}(x) = \emptyset$ occurs with probability at most $(1-p)^n \leq e^{-pn} = \delta$. Thus, with probability at least $1-\delta$ we have $\{X_i\}_{i\in[n]} \cap \overline{B}_{r(p)}(x) \neq \emptyset$ which implies $\rho(x, X_{\tau_1(x)}) \leq r(p)$, so $\mu(B_{\rho(x,X_{\tau_1(x)})}(x)) \leq \mu(B_{r(p)}(x)) \leq p = \log(1/\delta)/n$. □

Proof. [Proof of Theorem 4.1] Take $\zeta > 0$ and choose $x \in \mathcal{X}$ so that $f(x) > \sup_{z\in\mathcal{X}}\{f(z)\} - \zeta$. By Lemma 4.1 we have $\mu(B_{\rho(x,X_{\tau_1(x)})}(x)) \leq \log(2/\delta)/n$ with probability at least $1-\delta/2$. By Assumption 2.2 this implies

$$f(X_{\tau_1(x)}) \geq \sup_{z\in\mathcal{X}}\{f(z)\} - \omega \cdot \left(\frac{\log(2/\delta)}{n}\right)^{\lambda} - \zeta. \tag{4.1}$$

By Theorem 3.1 the following holds simultaneously for all $i \in [n]$ with probability at least $1 - \delta/2$,

$$\left|\hat{f}_{n,k}(X_i) - f(X_i)\right| \leq \xi(n,k,\delta) := \sigma\sqrt{\frac{2\log(4n/\delta)}{k}} + \omega \cdot \left(\frac{2k}{n}\right)^{\lambda}. \tag{4.2}$$

By the union bound we deduce that both (4.1) and (4.2) hold simultaneously with probability at least $1 - \delta$. Thus, we have

$$\begin{aligned} f(x_k(n)) &\geq \hat{f}_{n,k}(x_k(n)) - \xi(n,k,\delta) \\ &\geq \hat{f}_{n,k}(X_{\tau_1(x)}) - \xi(n,k,\delta) \\ &\geq f(X_{\tau_1(x)}) - 2\xi(n,k,\delta) \\ &\geq \sup_{z\in\mathcal{X}}\{f(z)\} - 2\xi(n,k,\delta) - \omega \cdot \left(\frac{\log(2/\delta)}{n}\right)^{\lambda} - \zeta. \end{aligned}$$

Letting $\zeta \to 0$ and applying continuity of measure we deduce

$$\mathcal{R}_f(x_k(n)) \leq 2\xi(n,k,\delta) + \omega \cdot \left(\frac{\log(2/\delta)}{n}\right)^{\lambda},$$

with probability at least $1 - \delta$. This completes the proof of the theorem. □

Theorem 4.1 is related to [27, Theorem 4.2] which gives a bound for an estimator of determining the maximum value of a function, rather than a regret bound for locating a point of near maximal value. In a recent work [26, Theorem 3] the present authors present a bound for determining the maximum value of a function in a flexible non-compact setting where Assumption 2.2 does not hold. We emphasise that the techniques in this paper do not extend to that setting, and obtaining a method for locating a point in the input space with near maximal value (rather than simply estimating the maximal value of the function) in the non-compact setting of [12,26] remains an important open question.

We now turn to the more challenging problem of robust optimisation.

5. Robust optimisation with k-NN

In this section we turn to the problem of robust optisation of the function f given a data sample $\mathcal{D} = \{(X_i, Y_i)\}_{i\in[n]}$, where we seek to maximise $\underline{f}^{\epsilon}(x)$, the infimum of the function value on a closed ball around x. Equivalently, our goal is to choose a point $x \in \mathcal{X}$ which minimises the adverserially robust regret $\mathcal{R}_f^{\epsilon}(x)$. We propose the following estimator:

$$x_k^{\epsilon}(n) := \operatorname*{argmax}_{X_i \,:\, i\in[n]} \left\{ \min_{j\in[n]} \left\{ \hat{f}_{n,k}(X_j) : X_j \in \overline{B}_{\epsilon}(X_i) \right\} \right\}.$$

We emphasise that this method is straightforward to implement given standard methods for computing k-nearest neighbours. Our method for selecting $x_k^\epsilon(n)$ satisfies the following high probability bound.

Theorem 5.1. *Suppose that Assumptions 2.1, 2.2 and 2.3 hold. Given any $n \in \mathbb{N}$, $\delta \in (0,1)$ and $k \in \{\lceil 4\log(6n/\delta)\rceil + 1, \cdots, \lfloor n/2 \rfloor\}$ with probability at least $1-\delta$ we have*

$$\mathcal{R}_f^\epsilon(x_k^\epsilon(n)) \leq 2\sigma\sqrt{\frac{2\log(12n/\delta)}{k}} + 7\omega \cdot \left(\frac{2k}{n}\right)^\lambda.$$

In particular, with $k = \Theta\left((\log(n/\delta))^{1/(2\lambda+1)} \cdot n^{2\lambda/(2\lambda+1)}\right)$ then with probability at least $1-\delta$ we have

$$\mathcal{R}_f^\epsilon(x_k^\epsilon(n)) \leq \mathcal{O}\left(\left(\frac{\log(n/\delta)}{n}\right)^{\lambda/(2\lambda+1)}\right).$$

In Section 6 we will see that this is the best possible rate, up to logarithmic and constant factors. The proof of Theorem 5.1 is more intricate than that of 4.1. The main difficulty lies in showing that when points X_i have a low value $\underline{f}^\epsilon(X_i)$ then there must be data points X_j in the ball $X_j \in \overline{B}_\epsilon(X_i)$ such that $f(X_j)$ is low. This requires a geometric analysis based upon Assumption 2.3, starting with the following lemma.

Lemma 5.1. *Suppose Assumption 2.3 holds. Given $p \in (0,1)$ and $x_0, x_1 \in \mathcal{X}$ with $\mu(B_{\rho(x_0,x_1)}(x_0)) > p$, there exists a point $x_* \in \mathcal{X}$ such that $B_{\rho(x_*,x_1)}(x_*) \subset B_{\rho(x_0,x_1)}(x_0)$, $\mu(B_{\rho(x_*,x_1)}(x_*)) \leq p$ and $\mu(\overline{B}_{\rho(x_*,x_1)}(x_*)) \geq p$.*

Proof. By Assumption 2.3 we obtain a $\rho(x_0,x_1)$-Lipschitz continuous path $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x_0$ and $\gamma(x_1) = x_1$. We define a pair of functions M, $\overline{M} : [0,1] \to [0, \mu(\overline{B}_{\rho(x_*,x_1)}(x_*))]$ by

$$M(t) := \mu(B_{\rho(\gamma(t),x_1)}(\gamma(t))), \qquad \overline{M}(t) := \mu(\overline{B}_{\rho(\gamma(t),x_1)}(\gamma(t))).$$

It may be shown that M is non-increasing, with $M(t) \leq \overline{M}(t)$ for all $t \in [0,1]$, M is right-continuous and $\overline{M}$ is left-continuous (see Lemma Appendix B.1). Now take $t_* := \sup\{t \in [0,1] : M(t) > p\}$. Note that $M(0) > p$ by construction, so $t_* \in (0,1]$ is well-defined. Moreover, $M(1) = 0$, so for $t \in (t_*, 1]$, $M(t) \leq p$ and so by the right-continuity of M we have $M(t_*) \leq p$. On the other hand, for $t \in [0, t_*)$ we have $\overline{M}(t) \geq M(t) > p$ and so by the left-continuity of $\overline{M}$, $\overline{M}(t_*) \geq p$. Thus, the lemma holds with $x_* = \gamma(t_*)$. □

Lemma 5.2. *Suppose Assumptions 2.2 and 2.3 hold and take $n \in \mathbb{N}$, $\delta \in (0,1)$ and $x_0 \in \mathcal{X}$ with $f(x_0) > \underline{f}^\epsilon(x_0) + \omega \cdot (\log(1/\delta)/n)^\lambda$. Then with probability at least $1-\delta$,*

$$\min_{i \in [n]} \left\{ f(X_i) : X_i \in \overline{B}_\epsilon(x_0) \right\} \leq \underline{f}^\epsilon(x_0) + 2\omega \cdot \left(\frac{\log(1/\delta)}{n}\right)^\lambda.$$

Proof. Let $p = \log(1/\delta)/n$. Take $\zeta \in (0, f(x_0) - \underline{f}^{\epsilon}(x_0) - \omega \cdot p^{\lambda})$ and choose $x_1 \in \overline{B}_{\epsilon}(x_0)$ with $f(x_1) < \underline{f}^{\epsilon}(x_0) + \zeta < f(x_0) - \omega \cdot p^{\lambda}$. By Assumption 2.2 this implies that $\mu(B_{\rho(x_0,x_1)}(x_0)) > p$. Hence, by Lemma 5.1 there exists a point $x_* \in \mathcal{X}$ such that $B_{\rho(x_*,x_1)}(x_*) \subset B_{\rho(x_0,x_1)}(x_0)$, $\mu(B_{\rho(x_*,x_1)}(x_*)) \leq p$ and $\mu(\overline{B}_{\rho(x_*,x_1)}(x_*)) \geq p$. By Lemma 4.1, with probability at least $1-\delta$ we have $\mu(B_{\rho(x_*,X_{\tau_1(x_*)})}(x_*)) \leq p$. This implies that $\rho(x_*, X_{\tau_1(x_*)}) \leq \rho(x_*, x_1)$, so $X_{\tau_1(x_*)} \in \overline{B}_{\rho(x_*,x_1)}(x_*) \subset \overline{B}_{\rho(x_0,x_1)}(x_0) \subset \overline{B}_{\epsilon}(x_0)$. Moreover, by Assumption 2.2 we have

$$\begin{aligned} f(X_{\tau_1(x_*)}) &\leq f(x_*) + \omega \cdot \mu(B_{\rho(x_*,X_{\tau_1(x_*)})}(x_*))^{\lambda} \\ &\leq f(x_1) + \omega \cdot \mu(B_{\rho(x_*,x_1)}(x_*))^{\lambda} + \omega \cdot p^{\lambda} \\ &\leq \underline{f}^{\epsilon}(x_0) + 2\omega \cdot p^{\lambda} + \zeta, \end{aligned}$$

with probability at least $1-\delta$. By taking $\zeta \to 0$, this completes the proof of the lemma. □

Lemma 5.3. *Suppose Assumptions 2.2 and 2.3 hold and take $n \in \mathbb{N}\backslash\{1\}$, $\delta \in (0,1)$. With probability at least $1-\delta$ the following holds for all $i \in [n]$,*

$$\min_{j\in[n]} \left\{ f(X_j) : X_j \in \overline{B}_{\epsilon}(X_i) \right\} \leq \underline{f}^{\epsilon}(X_i) + 2\omega \cdot \left(\frac{2\log(n/\delta)}{n} \right)^{\lambda}$$

Proof. Take $\zeta = 2\omega \cdot (2\log(n/\delta)/n)^{\lambda}$. For $q \in \{n-1, n\}$ we let $\boldsymbol{X}_{[q]} = \{X_i\}_{i\in[q]}$. We begin by fixing X_n and considering the conditional distribution over the i.i.d. sample $\boldsymbol{X}_{[n-1]}$. By applying Lemma 5.2 with X_n in place of x_0 and $\boldsymbol{X}_{[n-1]}$ in place of $\boldsymbol{X}_{[n-1]}$ we obtain

$$\mathbb{P}_{\boldsymbol{X}_{[n-1]}|X_n} \left[\min_{j\in[n]} \left\{ f(X_j) : X_j \in \overline{B}_{\epsilon}(X_n) \right\} > \underline{f}^{\epsilon}(X_n) + \zeta \right] \leq \frac{\delta}{n}.$$

By the law of total expectation we obtain

$$\mathbb{P}_{\boldsymbol{X}_{[n]}} \left[\min_{j\in[n]} \left\{ f(X_j) : X_j \in \overline{B}_{\epsilon}(X_n) \right\} > \underline{f}^{\epsilon}(X_n) + \zeta \right] \leq \frac{\delta}{n}. \tag{5.1}$$

By symmetry the same is true with X_i in place of X_n for any $i \in [n]$. The conclusion of the lemma follows with a union bound. □

Lemma 5.4. *Suppose that Assumptions 2.2 and 2.3 hold. Given any $x_0, x_1 \in \mathcal{X}$ we have*

$$\underline{f}^{\epsilon}(x_0) - \underline{f}^{\epsilon}(x_1) \leq \omega \cdot \mu \left(B_{\rho(x_0,x_1)}(x_0) \right)^{\lambda}.$$

Proof. Take $\zeta > 0$ and choose $x_2 \in \overline{B}_{\epsilon}(x_1)$ so that $f(x_2) < \underline{f}^{\epsilon}(x_1) + \zeta$. If $x_2 \in \overline{B}_{\epsilon}(x_0)$ then we have $\underline{f}^{\epsilon}(x_0) - \underline{f}^{\epsilon}(x_1) \leq \zeta$. If $x_2 \notin \overline{B}_{\epsilon}(x_0)$ we apply Assumption 2.3 to obtain a $\rho(x_0, x_2)$-Lipschitz continuous map $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x_0$ and

$\gamma(1) = x_2$. Take $x_3 = \gamma\left(\epsilon/\rho(x_0, x_2)\right)$, so $\rho(x_0, x_3) = \epsilon$ and $\rho(x_3, x_2) = \rho(x_0, x_2) - \epsilon$. Hence, $x_3 \in \overline{B}_\epsilon(x_0)$ and so

$$\begin{aligned}
\underline{f}^\epsilon(x_0) &\leq f(x_3) \\
&\leq f(x_2) + \omega \cdot \inf_{z \in \overline{B}_\epsilon(x_3)} \left\{ \mu\left(B_{\rho(x_2,x_3)}(z)\right)^\lambda \right\} \\
&\leq f(x_2) + \omega \cdot \mu\left(B_{\rho(x_0,x_1)}(x_0)\right)^\lambda \\
&\leq \underline{f}^\epsilon(x_1) + \omega \cdot \mu\left(B_{\rho(x_0,x_1)}(x_0)\right)^\lambda + \zeta,
\end{aligned}$$

where the penultimate inequality follows from the fact that $x_0 \in \overline{B}_\epsilon(x_3)$ and $\rho(x_2, x_3) = \rho(x_0, x_2) - \epsilon \leq \rho(x_0, x_1) + \rho(x_1, x_2) - \epsilon \leq \rho(x_0, x_1)$ as $\rho(x_1, x_2) \leq \epsilon$. Letting $\zeta \to 0$ completes the proof of the lemma. □

Proof. [Proof of Theorem 5.1] By Theorem 3.1 combined with Lemma 5.2 via a union bound we see that with probability at least $1 - 2\delta/3$ the following holds simultaneously for all $i \in [n]$,

$$\left| \min_{j \in [n]} \left\{ \hat{f}_{n,k}(X_j) : X_j \in \overline{B}_\epsilon(X_i) \right\} - \underline{f}^\epsilon(X_i) \right| \leq \underline{\xi}(n, k, \delta), \tag{5.2}$$

where

$$\underline{\xi}(n, k, \delta) := \sigma \sqrt{\frac{2\log(12n/\delta)}{k}} + 3\omega \cdot \left(\frac{2k}{n}\right)^\lambda.$$

Now take $\zeta > 0$ and choose $x_0 \in \mathcal{X}$ with $\underline{f}^\epsilon(x_0) > \sup_{z \in \mathcal{X}} \{\underline{f}^\epsilon(z)\} - \zeta$. By Lemma 4.1 we have $\mu(B_{\rho(x_0, X_{\tau_1(x_0)})}(x_0)) \leq \log(3/\delta)/n$ with probability at least $1 - \delta/3$. By Lemma 5.4 this implies that

$$\begin{aligned}
\underline{f}^\epsilon(X_{\tau_1(x_0)}) &\geq \underline{f}^\epsilon(x_0) - \omega \cdot \left(\frac{\log(3/\delta)}{n}\right)^\lambda \\
&\geq \sup_{z \in \mathcal{X}} \{\underline{f}^\epsilon(z)\} - \omega \cdot \left(\frac{\log(3/\delta)}{n}\right)^\lambda - \zeta.
\end{aligned}$$

By the union bound we can assume that (5.2) also holds, with total probability at least $1 - \delta$. Hence, with probability at least $1 - \delta$ we have

$$\begin{aligned}
\underline{f}^\epsilon(x_k^\epsilon(n)) &\geq \min_{j \in [n]} \left\{ \hat{f}_{n,k}(X_j) : X_j \in \overline{B}_\epsilon(x_k^\epsilon(n)) \right\} - \underline{\xi}(n, k, \delta) \\
&\geq \min_{j \in [n]} \left\{ \hat{f}_{n,k}(X_j) : X_j \in \overline{B}_\epsilon(X_{\tau_1(x_0)}) \right\} - \underline{\xi}(n, k, \delta) \\
&\geq \underline{f}^\epsilon(X_{\tau_1(x_0)}) - 2\underline{\xi}(n, k, \delta) \\
&\geq \sup_{z \in \mathcal{X}} \{\underline{f}^\epsilon(z)\} - 2\underline{\xi}(n, k, \delta) - \omega \cdot \left(\frac{\log(3/\delta)}{n}\right)^\lambda - \zeta.
\end{aligned}$$

Letting $\zeta \to 0$ and applying continuity of measure completes the proof of the theorem. □

6. A minimax lower bound

In this section we present a lower bound which demonstrates the optimality of the upper bounds in Theorems 4.1 and 5.1.

Theorem 6.1. *Let $\mathcal{X} = [0,1]$ with ρ the Euclidean metric, $\epsilon \in [0, 1/5]$, and ω, $\lambda > 0$. Given any method $\hat{x}_\epsilon(n)$ for locating the adverserially perturbed minimum, based upon a sample $\mathcal{D}$, there exists a distribution $\mathbb{P}$ on $\mathcal{X} \times \mathbb{R}$ with $f(x) = \mathbb{E}[Y|X = x]$ and satisfying Assumptions 2.1, 2.2, 2.3 and for $n \in \mathbb{N}$,*

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{R}_f^\epsilon\left(\hat{x}_\epsilon(n)\right)\right] \geq C_{\omega,\lambda} \cdot n^{-\frac{\lambda}{2\lambda+1}},$$

where the expectation is taken over samples $\mathcal{D} = \{(X_i, Y_i)\}_{i\in[n]}$ with $(X_i, Y_i) \sim \mathbb{P}$ i.i.d. and $C_{\omega,\lambda}$ is a constant determined by ω and λ.

The proof of the theorem is based on methods from non-parametric statistics [30] in which we introduce a pair of probability measures which are difficult to differentiate based on the sample, but are such that no selected point can perform well on both.

Given $q \in \mathbb{N}$ with $q \geq 5$ we shall construct a pair of probability measures $\mathbb{P}_0^q$ and $\mathbb{P}_1^q$ on $\mathcal{X} \times \mathbb{R}$ as follows. Let $\Delta_q := \omega \cdot (1/q)^\lambda$ and define a pair of functions $g_s^q : [0,1] \to [0,\infty)$ for $s \in \{0,1\}$ by

$$g_0^q(x) = \begin{cases} \Delta_q \cdot q \cdot x & \text{for } x \in \left[0, \frac{1}{q}\right] \\ \Delta_q & \text{for } x \in \left[\frac{1}{q}, \frac{1}{q} + 2\epsilon\right] \\ \Delta_q \cdot q \cdot \left(\left(\frac{2}{q} + 2\epsilon\right) - x\right) & \text{for } x \in \left[\frac{1}{q} + 2\epsilon, \frac{2}{q} + 2\epsilon\right] \\ 0 & \text{for } x \in \left[\frac{2}{q} + 2\epsilon, 1\right]. \end{cases}$$

$$g_1^q(x) = \begin{cases} 0 & \text{for } x \in \left[0, \frac{1}{q}\right] \\ \Delta_q \cdot q \cdot \left(x - \frac{1}{q}\right) & \text{for } x \in \left[\frac{1}{q}, \frac{2}{q}\right] \\ \Delta_q & \text{for } x \in \left[\frac{2}{q}, \frac{2}{q} + 2\epsilon\right] \\ \Delta_q \cdot q \cdot \left(\left(\frac{3}{q} + 2\epsilon\right) - x\right) & \text{for } x \in \left[\frac{2}{q} + 2\epsilon, \frac{3}{q} + 2\epsilon\right] \\ 0 & \text{for } x \in \left[\frac{3}{q} + 2\epsilon, 1\right]. \end{cases}$$

Let μ denote the Lebesgue measure restricted to $\mathcal{X} = [0,1]$. For $s \in \{0,1\}$ we let μ be the marginal distribution of $\mathbb{P}_s^q$ over $\mathcal{X}$ and for each $x \in \mathcal{X}$ we define the conditional distribution of Y given $X = x$ by $\mathbb{P}_s^q[Y|X = x] = \mathcal{N}\left(g_s^q(x), 1\right)$ i.e. a unit variance normal distribution with mean $g_s^q(x)$.

Lemma 6.1. *For both $s \in \{0,1\}$ the distribution $\mathbb{P}_s^q$ with $f = g_s^q$ satisfies Assumptions 2.1, 2.2 and 2.3.*

Proof. Assumption 2.1 follows from the fact that $\mathbb{P}_s^q[Y|X = x] = \mathcal{N}\left(g_s^q(x), 1\right)$ is Gaussian. Assumption 2.2 follows from Proposition 2.1. Assumption 2.3 follows

immediately from Proposition 2.2 (a) since $\mathcal{X} = [0,1]$ with the Euclidean metric is a convex subset of a normed vector space. □

We utlised the following standard lemma.

Lemma 6.2 (Tsybakov [30]). *Suppose that there are probability distributions $\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1$ on a measurable space $\mathcal{Z}$, with $\tilde{\mathbb{P}}_1$ absolutely continuous with respect to $\tilde{\mathbb{P}}_0$. Then for any measureable function $\tau : \mathcal{Z} \to \{0,1\}$ we have*

$$\tilde{\mathbb{P}}_0[\tau(Z)=1] + \tilde{\mathbb{P}}_1[\tau(Z)=0] \geq \frac{1}{2} \cdot \exp\left(-KL\left(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1\right)\right).$$

We shall apply Lemma 6.2 with respect to the product measures $(\mathbb{P}_s^q)^{\otimes n}$ for $s \in \{0,1\}$ on $\mathcal{Z} = (\mathcal{X} \times \mathbb{R})^n$.

Lemma 6.3. $KL\left((\mathbb{P}_0^q)^{\otimes n}, (\mathbb{P}_1^q)^{\otimes n}\right) \leq \frac{2n}{q} \cdot (\Delta_q)^2$.

Proof. Note that g_0^q and g_1^q only differ for $x \in [0, 2/q]$ and $x \in [1/q + 2\epsilon, 3/q + 2\epsilon]$ and $\|g_0^q - g_1^q\|_\infty = \Delta_q$. Hence, we have

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_0^q, \mathbb{P}_1^q\right) &= \int \mathrm{KL}\left(\mathcal{N}\left(g_0^q(x), 1\right), \mathcal{N}\left(g_1^q(x), 1\right)\right) d\mu(x) \\
&= \frac{1}{2} \int \left(g_0^q(x) - g_1^q(x)\right)^2 d\mu(x) \\
&\leq \frac{1}{2} \cdot \frac{4}{q} \cdot (\Delta_q)^2.
\end{aligned}$$

The lemma follows since $\mathrm{KL}\left((\mathbb{P}_0^q)^{\otimes n}, (\mathbb{P}_1^q)^{\otimes n}\right) = n \cdot \mathrm{KL}\left(\mathbb{P}_0^q, \mathbb{P}_1^q\right)$. □

Proof. [Proof of Theorem 6.1] To complete the proof we take $q = \lceil (2n)^{\frac{1}{2\lambda+1}} \rceil$ so by Lemma 6.3 we have $\mathrm{KL}\left((\mathbb{P}_0^q)^{\otimes n}, (\mathbb{P}_1^q)^{\otimes n}\right) \leq \omega^2$. Suppose we have an estimator $\hat{x}_\epsilon(n)$ which is measurable with respect to the sample $\mathcal{D}$. We consider the hypothesis test $\tau : (\mathcal{X} \times \mathbb{R})^n \to \{0,1\}$ by $\tau(\mathcal{D}) = 0$ if $\hat{x}_\epsilon(n) \leq 3/(2q) + \epsilon$ and $\tau(\mathcal{D}) = 1$ otherwise. By Lemma 6.2 we have

$$\left(\mathbb{P}_0^q\right)^{\otimes n}\left[\hat{x}_\epsilon(n) > \frac{3}{2q} + \epsilon\right] + \left(\mathbb{P}_1^q\right)^{\otimes n}\left[\hat{x}_\epsilon(n) \leq \frac{3}{2q} + \epsilon\right] \geq \frac{1}{2e^{\omega^2}}. \tag{6.1}$$

Now it follows from the construction of g_0^q that if $\hat{x}_\epsilon(n) > 3/(2q) + \epsilon$ then $\mathcal{R}^\epsilon_{g_0^q}(\hat{x}_\epsilon(n)) \geq \Delta_q/2$. On the other hand, if $\hat{x}_\epsilon(n) \leq 3/(2q) + \epsilon$ then $\mathcal{R}^\epsilon_{g_1^q}(\hat{x}_\epsilon(n)) \geq \Delta_q/2$. Hence, by (6.1) we have

$$\min_{s \in \{0,1\}} \left\{ \left(\mathbb{P}_s^q\right)^{\otimes n} \left[\mathcal{R}^\epsilon_{g_s^q}(\hat{x}_\epsilon(n))\right] \right\} \geq \frac{\Delta_q}{8e^{\omega^2}} \geq \frac{\omega}{2^{3+2\lambda} e^{\omega^2}} \cdot n^{-\frac{\lambda}{2\lambda+1}}.$$

By Lemma 6.1 this completes the proof of the theorem. □

7. Discussion

We have considered the problem of robust randomised black box optimisation of a function on an arbitrary metric space, given a data set of noisy function evaluations. We have presented a conceptually simple method based on k-nearest neighbour regression for finding the adverserially robust optimum, with a high probability performance upper bound (Theorem 5.1). Moreover, we have shown that the performance of our method is optimal up to logarithmic and constant factors (Theorem 6.1). Whilst the present work is theoretical in nature, our conceptually simple and minimax optimal method displays potential for applications in hyper-parameter tuning for complex models on large data sets.

Acknowledgements

This work is funded by EPSRC under Fellowship grant EP/P004245/1.

References

[1] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263, 2018.

[2] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[3] Aharon Ben-Tal and Arkadi Nemirovski. Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002.

[4] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[5] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[6] Dimitris Bertsimas, Omid Nohadani, and Kwong Meng Teo. Nonconvex robust optimization for problems with constraints. *INFORMS journal on computing*, 22(1):44–58, 2010.

[7] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

[8] Ilija Bogunovic, Slobodan Mitrović, Jonathan Scarlett, and Volkan Cevher. Robust submodular maximization: A non-uniform partitioning approach. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 508–516. JMLR. org, 2017.

[9] Ilija Bogunovic, Jonathan Scarlett, Stefanie Jegelka, and Volkan Cevher. Adversarially robust optimization with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5765–5775, 2018.

[10] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

[11] Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18:227:1–227:16, 2017.

[12] Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general fi-

nite dimensional spaces with the k nearest neighbour rule. *The Annals of Statistics*, 44(3):982–1009, 06 2016.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[14] Lee-Ad Gottlieb and Aryeh Kontorovich. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

[15] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 370–378, 2014.

[16] W Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. *arXiv preprint arXiv:1906.03291*, 2019.

[17] Heinrich Jiang. Non-asymptotic uniform rates of consistency for k-nn regression. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI, 2019.

[18] Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In *International Conference on Machine Learning*, pages 892–900, 2014.

[19] Aryeh Kontorovich and Roi Weiss. A bayes consistent 1-nn classifier. In *Artificial Intelligence and Statistics*, pages 480–488, 2015.

[20] Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

[21] Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492, 2018.

[22] Hrushikesh N Mhaskar and Tomaso Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[23] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

[24] Michael A Osborne, Roman Garnett, and Stephen J Roberts. Gaussian processes for global optimization. In *Bayesian Optimization*, 2009.

[25] Min Xu Qin Fang and Yiming Ying. Faster convergence of a randomized coordinate descent method for linearly constrained optimization problems. *Analysis and Applications*, 16(05):741–755, 2018.

[26] Henry W. J. Reeve and Ata Kabán. Classification with unknown class-conditional label noise on non-compact feature spaces. In *Proceedings of the 32nd Conference on Learning Theory.*, volume 99 of *Proceedings of Machine Learning Research*, Pheonix, Arizona, 25–28 Jul 2019. PMLR.

[27] Henry W. J. Reeve and Ata Kabán. Fast rates for a knn classifier robust to unknown asymmetric label noise. In *Proceedings of the 36th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, Long Beach, California, 10–15 Jul 2019. PMLR.

[28] Christoph Schwab and Jakob Zech. Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in uq. *Analysis and Applications*, 17(01):19–55, 2019.

[29] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[30] Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

[31] Ding-Xuan Zhou. Deep distributed convolutional neural networks: Universality. *Analysis and Applications*, 16(06):895–919, 2018.

Appendix A. Proof of Theorem 3.1

We begin by proving several supporting lemmas. Throughout the proof we let $\boldsymbol{X}_{[n]}$ denote $\{X_i\}_{i\in[n]}$ and let $r_{n,k}(x)$ denote the distance to the k-th nearest neighbour i.e. $r_{n,k}(x) := \rho(x, X_{\tau_k(x)})$. The following lemma is a consequence of the multiplicative Chernoff bound [23, Theorem 4.5]. See [27, Lemma 3.2] for a proof.

Lemma Appendix A.1. *Given $x \in \mathcal{X}$, $\delta \in (0,1)$ and $k \in \{\lceil 4\log(1/\delta)\rceil, \cdots, \lfloor n/2\rfloor\}$. Then with probability at least $1-\delta$ we have $\mu\left(B_{r_{n,k}(x)}(x)\right) \leq (2k)/n$.*

We now deduce Corollary Appendix A.1.

Corollary Appendix A.1. *Given $\delta \in (0,1)$ and $k \in \{\lceil 4\log(n/\delta)\rceil + 1, \cdots, \lfloor n/2\rfloor\}$. Then with probability at least $1-\delta$, we have $\mu\left(B_{r_{n,k}(X_i)}(X_i)\right) \leq (2k)/n$ simultaneously for all $i \in [n]$.*

Proof. We begin by considering the conditional distribution over $\boldsymbol{X}_{[n-1]} = \{X_i\}_{i\in[n-1]}$ given X_n. We apply Lemma Appendix A.1 with X_n fixed, considering $\{X_i\}_{i\in[n-1]}$ in place of $\{X_i\}_{i\in[n]}$ as follows,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{X}_{[n-1]}|X_n}\left[\mu\left(B_{r_{n,k}(X_n)}(X_n)\right) > \frac{2k}{n}\right] \\
&\leq \mathbb{P}_{\boldsymbol{X}_{[n-1]}|X_n}\left[\mu\left(B_{r_{n-1,k-1}(X_n)}(X_n)\right) > \frac{2k}{n}\right] \\
&\leq \mathbb{P}_{\boldsymbol{X}_{[n-1]}|X_n}\left[\mu\left(B_{r_{n-1,k-1}(X_n)}(X_n)\right) > \frac{2(k-1)}{n-1}\right] \leq \frac{\delta}{n}.
\end{aligned}$$

By the law of total expectation we have

$$\mathbb{P}_{\boldsymbol{X}_{[n]}}\left[\mu\left(B_{r_{n,k}(X_n)}(X_n)\right) > \frac{2k}{n}\right] \leq \frac{\delta}{n}.$$

By symmetry the same is true with any X_i in place of X_n. Combining the above with the union bound completes the proof of the lemma. □

We also require the following consequence of Hoeffing's inequality.

Lemma Appendix A.2. *Given $\delta \in (0,1)$ and $k \in [n]$, the following holds simultaneously for all $i \in [n]$, with probability at least $1-\delta$,*

$$\left|\hat{f}_{n,k}\left(X_i\right) - \frac{1}{k}\sum_{q\in[k]} f\left(X_{\tau_q(X_i)}\right)\right| \leq \sigma\sqrt{\frac{2\log(2n/\delta)}{k}}.$$

Proof. Take $t = \sigma\sqrt{2\log(2n/\delta)/k}$. For each $i \in [n]$ the following holds by the

Hoeffding bound,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{Y}|\boldsymbol{X}_{[n]}}\left[\left|\hat{f}_{n,k}\left(X_i\right)-\frac{1}{k}\sum_{q\in[k]}f\left(X_{\tau_q(X_i)}\right)\right|>t\right]\\
&\leq \mathbb{P}_{\boldsymbol{Y}|\boldsymbol{X}_{[n]}}\left[\left|\frac{1}{k}\sum_{q\in[k]}(Y_{\tau_q(X_i)}-f\left(X_{\tau_q(X_i)}\right))\right|>t\right]\\
&\leq 2\exp\left(-\frac{k\cdot t^2}{2\sigma^2}\right)\leq\frac{\delta}{n}.
\end{aligned}$$

By the union bound over $i\in[n]$, the lemma follows. □

Proof. [Proof of Theorem 3.1] By Lemmas Appendix A.1, Appendix A.2 combined with the union bound, we have both

$$\left|\hat{f}_{n,k}\left(X_i\right)-\frac{1}{k}\sum_{q\in[k]}f\left(X_{\tau_q(X_i)}\right)\right|\leq\sigma\sqrt{\frac{2\log(4n/\delta)}{k}},\tag{A.1}$$

and $\mu\left(B_{r_{n,k}(X_i)}(X_i)\right)\leq(2k)/n$ simultaneously for all $i\in[n]$, with probability at least $1-\delta$. To complete the proof it suffices to observe that by Assumption 2.2, $\mu\left(B_{r_{n,k}(X_i)}(X_i)\right)\leq(2k)/n$ implies that for each $q\in[k]$,

$$\begin{aligned}
\left|f\left(X_{\tau_q(X_i)}\right)-f(X_i)\right|&\leq\omega\cdot\mu\left(B_{\rho(X_i,X_{\tau_q(X_i)})}(X_i)\right)^{\lambda}\\
&\leq\omega\cdot\mu\left(B_{\rho(X_i,X_{\tau_k(X_i)})}(X_i)\right)^{\lambda}\\
&=\omega\cdot\mu\left(B_{r_{n,k}(X_i)}(X_i)\right)^{\lambda}\leq\omega\cdot\left(\frac{2k}{n}\right)^{\lambda}.
\end{aligned}$$

In conjunction with (A.1) this completes the proof of Theorem 3.1. □

Appendix B. Technical lemma for the proof of Theorem 5.1

The following technical lemma underpins the proof of Lemma 5.1 which is required for the proof of Theorem 5.1.

Lemma Appendix B.1. *Suppose that Assumption 2.3 holds and take $x_0,x_1\in\mathcal{X}$. Let $\gamma:[0,1]\to\mathcal{X}$ be a $\rho(x_0,x_1)$-Lipschitz continuous path and define a pair of functions M, $\overline{M}:[0,1]\to[0,\mu(\overline{B}_{\rho(x_*,x_1)}(x_*))]$ by $M(t):=\mu(B_{\rho(\gamma(t),x_1)}(\gamma(t)))$ and $\overline{M}(t):=\mu(\overline{B}_{\rho(\gamma(t),x_1)}(\gamma(t)))$. Then the following holds.*

(1) M, $\overline{M}$ are non-increasing on $[0,1]$;
(2) $M(t)\leq\overline{M}(t)$ for all $t\in[0,1]$;
(3) M is right-continuous on $[0,1]$;
(4) $\overline{M}$ is left-continuous on $[0,1]$.

Proof. We begin with two observations. Firstly, (I) Given $t_a, t_b \in (0,1)$ we have $\rho(\gamma(t_a), \gamma(t_b)) = |t_a - t_b| \cdot \rho(x_0, x_1)$. Indeed, by the $\rho(x_0, x_1)$-Lipschitz property we have $\rho(\gamma(t_a), \gamma(t_b)) \leq |t_a - t_b| \cdot \rho(x_0, x_1)$. Now without loss of generality we may assume $t_a \leq t_b$ and by the triangle inequality

$$\begin{aligned}\rho(x_0, x_1) &\leq \rho(x_0, \gamma(t_a)) + \rho(\gamma(t_a), \gamma(t_b)) + \rho(\gamma(t_b), x_1) \\ &\leq t_a \cdot \rho(x_0, x_1) + \rho(\gamma(t_a), \gamma(t_b)) + (1 - t_b) \cdot \rho(x_0, x_1).\end{aligned}$$

Rearranging yields (I). Our second observation is that (II) $\{B_{\rho(\gamma(t), x_1)}(\gamma(t))\}_{t \in [0,1]}$ forms a decreasing nested family i.e. given $t_a < t_b$ we have $B_{\rho(\gamma(t_b), x_1)}(\gamma(t_b)) \subset B_{\rho(\gamma(t_a), x_1)}(\gamma(t_a))$. Indeed, if $\rho(z, \gamma(t_b)) < \rho(\gamma(t_b), x_1)$ then by observation (I) we have

$$\begin{aligned}\rho(z, \gamma(t_a)) &\leq \rho(z, \gamma(t_b)) + \rho(\gamma(t_b), \gamma(t_a)) \\ &< \rho(\gamma(t_b), x_1) + \rho(\gamma(t_b), \gamma(t_a)) \\ &= (1 - t_b)\rho(x_0, x_1) + (t_b - t_a)\rho(x_0, x_1) \\ &= \rho(\gamma(t_a), x_1).\end{aligned}$$

Lemma Appendix B.1 (1) now follows immediately from Observation (II). Lemma Appendix B.1 (2) is immediate from the definitions of M and $\overline{M}$.

To prove that M is right-continuous we take $t_\infty \in [0,1)$ and take a decreasing sequence $(t_q)_{q \in \mathbb{N}} \subset (t_\infty, 1)$ with $\lim_{q \to \infty} t_q = t_\infty$. By Lemma Appendix B.1 (1) we have $M(t_\infty) \geq \limsup_{q \to \infty} \{M(t_q)\}$. To prove Lemma Appendix B.1 (3) we must show that $M(t_\infty) \leq \liminf_{q \to \infty} \{M(t_q)\}$. We shall first prove the claim that $B_{\rho(\gamma(t_\infty), x_1)}(\gamma(t_\infty)) \subseteq \bigcup_{q \in \mathbb{N}} B_{\rho(\gamma(t_q), x_1)}(\gamma(t_q))$. To prove the claim take $z \in B_{\rho(\gamma(t_\infty), x_1)}(\gamma(t_\infty))$, so $\rho(z, \gamma(t_\infty)) < \rho(\gamma(t_\infty), x_1) = (1 - t_\infty) \cdot \rho(x_0, x_1)$, by observation (I). Hence, we may choose q sufficiently large that $\rho(z, \gamma(t_\infty)) < (1 - 2t_q + t_\infty) \cdot \rho(x_0, x_1)$, so we have

$$\begin{aligned}\rho(z, \gamma(t_q)) &\leq \rho(z, \gamma(t_\infty)) + \rho(\gamma(t_q), \gamma(t_\infty)) \\ &< (1 - 2t_q + t_\infty) \cdot \rho(x_0, x_1) + (t_q - t_\infty) \cdot \rho(x_0, x_1) \\ &= (1 - t_q) \cdot \rho(x_0, x_1) = \rho(\gamma(t_q), x_1),\end{aligned}$$

and so $z \in B_{\rho(\gamma(t_q), x_1)}(\gamma(t_q))$. This proves the claim that $B_{\rho(\gamma(t_\infty), x_1)}(\gamma(t_\infty)) \subseteq \bigcup_{q \in \mathbb{N}} B_{\rho(\gamma(t_q), x_1)}(\gamma(t_q))$. Moreover, by observation (II), $\left\{B_{\rho(\gamma(t_q), x_1)}(\gamma(t_q))\right\}_{q \in \mathbb{N}}$ is an increasing sequence of nested sets so by continuity of measure from below we have

$$\begin{aligned}M(t_\infty) &= \mu\left(B_{\rho(\gamma(t_\infty), x_1)}(\gamma(t_\infty))\right) \\ &\leq \mu\left(\bigcup_{q \in \mathbb{N}} B_{\rho(\gamma(t_q), x_1)}(\gamma(t_q))\right) \\ &= \lim_{q \to \infty} \mu\left(B_{\rho(\gamma(t_q), x_1)}(\gamma(t_q))\right) = \lim_{q \to \infty} \{M(t_q)\}.\end{aligned}$$

This completes the proof of Lemma Appendix B.1 (3).

The proof of Lemma Appendix B.1 (4) is similar. Take $t_\infty \in (0,1]$ and take an increasing sequence $(t_q)_{q\in\mathbb{N}} \subset (0,t_\infty)$ with $\lim_{q\to\infty} t_q = t_\infty$. By Lemma Appendix B.1 (1) we have $\overline{M}(t_\infty) \leq \liminf_{q\to\infty}\{M(t_q)\}$. To prove Lemma Appendix B.1 (4) we must show that $M(t_\infty) \geq \limsup_{q\to\infty}\{M(t_q)\}$. We first prove the claim that $\bigcap_{q\in\mathbb{N}} \overline{B}_{\rho(\gamma(t_q),x_1)}(\gamma(t_q)) \subset \overline{B}_{\rho(\gamma(t_\infty),x_1)}(\gamma(t_\infty))$. Indeed, given $z \in \bigcap_{q\in\mathbb{N}} \overline{B}_{\rho(\gamma(t_q),x_1)}(\gamma(t_q))$ for each $q \in \mathbb{N}$ we have

$$\begin{aligned}\rho(z,\gamma(t_\infty)) &\leq \rho(z,\gamma(t_q)) + \rho(\gamma(t_q),\gamma(t_\infty)) \\ &< \rho(\gamma(t_q),x_1) + \rho(\gamma(t_q),\gamma(t_\infty)) \\ &= ((1-t_q)+(t_\infty-t_q))\,\rho(x_0,x_1) \\ &= \rho(\gamma(t_\infty),x_1) + 2(t_\infty-t_q)\cdot\rho(x_0,x_1).\end{aligned}$$

Letting $q\to\infty$ we deduce that $z \in \overline{B}_{\rho(\gamma(t_\infty),x_1)}(\gamma(t_\infty))$. This completes the proof of the claim. Hence, by continuity of measure from above combined with observation (1) we have

$$\begin{aligned}\overline{M}(t_\infty) &= \mu\left(\overline{B}_{\rho(\gamma(t_\infty),x_1)}(\gamma(t_\infty))\right) \\ &\geq \mu\left(\bigcap_{q\in\mathbb{N}} \overline{B}_{\rho(\gamma(t_q),x_1)}(\gamma(t_q))\right) \\ &= \lim_{q\to\infty} \mu\left(\overline{B}_{\rho(\gamma(t_q),x_1)}(\gamma(t_q))\right) = \lim_{q\to\infty}\left\{\overline{M}(t_q)\right\}.\end{aligned}$$

This completes the proof of Lemma Appendix B.1 (4). □

Appendix C. Proof of Proposition 2.1

Proof. [Proof of Proposition 2.1] Take any $\epsilon > 0$ and choose $x_0, x_1 \in \mathcal{X}$. First suppose $\rho(x_0,x_1) < 1$. Then for any $z \in \mathcal{X}$ we have $\mu\left(B_{\rho(x_0,x_1)}(z)\right) \geq c_d \cdot \rho(x_0,x_1)^d$ and so

$$|f(x_0) - f(x_1)| \leq C_\alpha \cdot \rho(x_0,x_1)^\alpha \leq C_\alpha \cdot c_d^{-\alpha/d} \cdot \mu\left(B_{\rho(x_0,x_1)}(z)\right)^{\alpha/d}.$$

On the other hand, if $\rho(x_0,x_1) \geq 1$ then we have $\mu\left(B_{\rho(x_0,x_1)}(z)\right) \geq \mu\left(B_1(z)\right) \geq c_d$ and $f(x_0), f(x_1) \in [0,1]$ so

$$|f(x_0) - f(x_1)| \leq 1 \leq c_d^{-\alpha/d} \cdot \mu\left(B_{\rho(x_0,x_1)}(z)\right)^{\alpha/d}.$$

By considering these two cases, this completes the proof of the proposition. □